Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Arthroplasty. Author manuscript; available in PMC 2024 October 01.

Published in final edited form as:

J Arthroplasty. 2023 October ; 38(10): 2081–2084. doi:10.1016/j.arth.2022.10.031.

External Validation of Natural Language Processing Algorithms to Extract Common Data Elements in THA Operative Notes

Cody C. Wyles, MD[a,b,*], **Sunyang Fu, PhD**[c], **Susan L. Odum, PhD**[d], **Taylor Rowe, BS**[d], **Nahir A. Habet, MS**[d], **Daniel J. Berry, MD**[a], **David G. Lewallen, MD**[a], **Hilal Maradit-Kremers, MD**[b,c], **Sunghwan Sohn, PhD**[b,c], **Bryan D. Springer, MD**[d]

[a] Department of Orthopedic Surgery, Mayo Clinic, Rochester, Minnesota

[b] Orthopedic Surgery Artificial Intelligence Laboratory, Mayo Clinic, Rochester, Minnesota

[c] Department of AI and Informatics, Mayo Clinic, Rochester, Minnesota

[d] OrthoCarolina Research Institute, Charlotte, North Carolina

Abstract

Background: Natural language processing (NLP) systems are distinctive in their ability to extract critical information from raw text in electronic health records (EHR). We previously developed three algorithms for total hip arthroplasty (THA) operative notes with rules aimed at capturing (1) operative approach, (2) fixation method, and (3) bearing surface using inputs from a single institution. The purpose of this study was to externally validate and improve these algorithms as a prerequisite for broader adoption in automated registry data curation.

Methods: The previous NLP algorithms developed at Mayo Clinic were deployed and refined on EHRs from OrthoCarolina, evaluating 39 randomly selected primary THA operative reports from 2018 to 2021. Operative reports were available only in PDF format, requiring conversion to “readable” text with Adobe software. Accuracy statistics were calculated against manual chart review.

Results: The operative approach, fixation technique, and bearing surface algorithms all demonstrated perfect accuracy of 100%. By comparison, validated performance at the developing center yielded an accuracy of 99.2% for operative approach, 90.7% for fixation technique, and 95.8% for bearing surface.

Conclusion: NLP algorithms applied to data from an external center demonstrated excellent accuracy in delineating common elements in THA operative notes. Notably, the algorithms had no functional problems evaluating scanned PDFs that were converted to “readable” text by common software. Taken together, these findings provide promise for NLP applied to scanned PDFs as a source to develop large registries by reliably extracting data of interest from very large unstructured data sets in an expeditious and cost-effective manner.

* Address correspondence to: Cody C. Wyles, MD, Division of Hip and Knee Reconstruction, Department of Orthopedic Surgery, Mayo Clinic, 200 1st Street SW, Rochester, MN, 55905.

Investigation performed jointly at the Mayo Clinic in Rochester, MN and the OrthoCarolina Research Institute in Charlotte, NC.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Keywords

natural language processing; registry science; electronic medical records; artificial intelligence; total hip arthroplasty

Natural language processing (NLP) tools are distinctive in their ability to extract critical information from raw text in electronic health records (EHR) [1,2]. These tools have powerful implications for arthroplasty registries, potentially enabling (1) creation of local registries *de novo*, (2) augmentation of established registries, and (3) contribution to national registries such as the American Joint Replacement Registry (AJRR).

We previously developed three algorithms for total hip arthroplasty (THA) operative notes with rules aimed at capturing (1) operative approach, (2) fixation method, and (3) bearing surface [3]. These algorithms performed well in the tertiary care environment where they were trained and in an initial external validation test set from a community health system in the same enterprise. The next critical step was to test and improve the algorithms in a new center with different medical record systems and practice setup. Thus, the purpose of this study was to externally validate and improve these algorithms as a prerequisite for broader adoption in automated registry data curation.

Methods

Creation details of the NLP algorithms for surgical approach, bearing surface, and fixation method can be found in the original publication [3]. In brief, the algorithms were trained on 300 randomly selected primary THA operative notes from 29 different surgeons from 2000 to 2015 at the Mayo Clinic in Rochester, Minnesota (MCR). Three rule-based algorithms were created through an iterative process by content experts in data science and orthopaedic surgery and were hosted by an open-source NLP software MedTaggerIE. The algorithms aimed to use free text in operative notes to classify (1) surgical approach (posterior, lateral, direct anterior), (2) fixation method (uncemented, cemented, hybrid, reverse hybrid), and (3) bearing surface (ceramic-on-polyethylene, metal-on-polyethylene, ceramic-on-ceramic, metal-on-metal). Following initial algorithm development, they were tested in the Mayo Clinic Health System (MCHS) on 180 randomly selected operative notes from the same period as a means of external validation within a different community practice setting, but part of the same enterprise.

For this external validation effort at another tertiary academic center in another part of the country, we evaluated 39 randomly selected primary THA operative reports from two surgeons performed at the OrthoCarolina Hip and Knee Center in Charlotte, North Carolina (OC) from 2018 to 2021. One surgeon routinely performs THA with a posterior approach and the other performs a direct anterior approach. Both surgeons use a mix of ceramic-on-polyethylene and metal-on-polyethylene bearings, and both use a mix of uncemented fixation and hybrid fixation.

Unlike the investigations performed at MCR and MCHS, there were challenges at OC linking the MedTaggerIE software with the local EHR. This represented an opportunity to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

test a new means of information extraction. Operative reports were obtained as scanned and nonsearchable portable document formats (PDFs) and subsequently converted to “readable” text with standard issue Adobe (San Jose, California) Acrobat software and PDFMiner (a Python-based text extraction tool for PDF documents). These converted documents were then fed to MedTaggerIE by local OC team members for execution of the three algorithms. The workflow plan was to handle initial discrepancies in output through simultaneous review by a data scientist and orthopaedic surgeon, followed by iterative improvement of the source code and rerunning of the algorithms until satisfactory performance was achieved. Accuracy was calculated using manual chart review as the gold standard and was compared to internally validated performance at MCR and MCHS.

Results

During the first round of evaluation, two small sources of systematic error were identified and used to iteratively improve the algorithms. Through simultaneous review by an orthopaedic surgeon and data scientist, this first round of iterative improvement took less than 1 hour. Both sources of error involved updating the rules-based logic to be more inclusive to capture dictation styles of the local surgeons (Fig. 1). Notably, no error was generated by the workflow of (1) scanning PDFs from the EHR, (2) converting PDFs to readable text via Adobe Acrobat, and (3) running the converted operative reports through MedTaggerIE (Fig. 2).

Once implemented, the operative approach, fixation technique, and bearing surface algorithms all demonstrated perfect accuracy of 100%. By comparison, original performance at MCR yielded an accuracy of 99.2% for operative approach, 90.7% for fixation technique, and 95.8% for bearing surface and performance at MCHS yielded an accuracy of 94.4% for operative approach, 95.6% for fixation technique, and 98.0% for bearing surface (Table 1).

Discussion

NLP-enabled algorithms are a promising alternative to the current gold standard of manual chart review for the extraction and evaluation of large data sets in orthopaedics [2]. We previously developed algorithms at a tertiary arthroplasty center to automatically extract categorical features from THA operative reports including surgical approach, fixation method, and bearing surface [3]. These algorithms performed well locally and during external validation in a community health system linked to the same academic institution. In this study, we expanded on this effort and tested the algorithms in a new tertiary center, with a mix of challenges and successes. First, there was difficulty given local information technology capacity to interface the NLP software with the EHR. However, this created an opportunity to test a new strategy of downloading scanned operative note PDFs, converting them to readable text with simple Adobe Acrobat PDFMiner software, and running these documents through the NLP software. This proved successful. For those that do not already have the functionality, an Adobe Acrobat subscription that includes PDFMiner is less than $180/y. Furthermore, we were able to improve generalizability of the algorithms through a round of iterative improvement to account for local dictation nuance, ultimately yielding perfect performance on the test set.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Creating local registries is extremely difficult and requires enormous human resource and information technology investment. It is also challenging to augment existing registries with additional datapoints that were not part of the initial charter for prospective collection. NLP is one strategy to ease this burden with technological advancement. Algorithms designed to automatically abstract data from unstructured data in EHRs does require initial frontloaded effort in concert between data scientists and surgeon content experts. However, once refined through iterative improvement, these algorithms can autonomously abstract datapoints of interest on vast datasets [3–7]. Tools of this nature are published with a near universal sentiment in the limitations and conclusions that “external validation is required” or “further research is needed”. Too often, followup reports do not ensue, missing a critical opportunity to substantiate and improve valuable models. This study demonstrates the importance of external validation both to document external validity, but also to improve a model by adding features that enable it to achieve broader generalizability and overcome challenges of data shift (Fig. 1).

Another novel aspect of this study, arrived upon by happenstance, was learning that NLP software does not have to directly interface with the EHR to yield results. When difficulties arose in our attempt to do this at OC, we attempted to circumvent the problem through a seemingly clumsy workflow of obtaining PDFs, converting them to readable text, and then running those documents through the NLP software. Contrary to our expectation, the workflow was not particularly burdensome and showed a pathway to success that can be replicated by any center no matter how resource constrained. Perhaps this workflow will enable broader participation in large registries such as AJRR by lowering the burden of contribution necessary to abstract meaningful datasets. It may be as simple as sending PDFs of consult notes and operative reports to be dissected by centrally located NLP-capable teams, which could expand the horizon for the breadth and depth of information attainable by AJRR (Fig. 2).

This study should be interpreted considering the following limitations. First, the sample size of this second stage external validation study is small and only evaluated 2 surgeons, whereas the original study evaluated 29. Thus, the perfect performance of the algorithms after one round of iterative improvement to account for local dictation styles may not represent the level of difficulty in adapting the algorithms to other situations. Secondly, these algorithms address relatively simple features to extract from operative notes. For NLP to broadly potentiate detailed registry construction from clinical and operative notes, more sophisticated algorithms will need to be created that may not match the performance of the algorithms presented here. Nevertheless, these results are promising and the demonstration of iterative improvement through multiple rounds of external validation should become common place in orthopaedic research generally and AI-driven research in particular.

Conclusion

NLP algorithms applied to data from an external center demonstrated perfect accuracy in delineating common elements in THA operative notes following one round of iterative improvement. Notably, the algorithms had no functional problem evaluating scanned PDFs that were converted to “readable” text by common software. Taken together, these findings

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

provide promise for NLP applied to scanned PDFs as a source to develop large registries by reliably extracting data of interest from very large unstructured data sets in an expeditious and cost-effective manner. Encouragingly, this technology may lower the barrier for many institutions to participate in regional and national registry efforts and potentiate more detailed data abstraction from various EHRs.

Funding:

This work was funded in part by National Institutes of Health (NIH) [grant numbers R01AR73147 and P30AR76312] and supported by the American Joint Replacement Research-Collaborative (AJRR-C).

References

[1]. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 2005;12: 448–57. [PubMed: 15802475]

[2]. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011;306:848–55. [PubMed: 21862746]

[3]. Wyles CC, Tibbo ME, Fu S, Wang Y, Sohn S, Kremers WK, et al. Use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty. J Bone Joint Surg Am 2019;101:1931–8. [PubMed: 31567670]

[4]. Sagheb E, Ramazanian T, Tafti AP, Fu S, Kremers WK, Berry DJ, et al. Use of natural language processing algorithms to identify common data elements in operative notes for knee arthroplasty. J Arthroplasty 2021;36:922–6. [PubMed: 33051119]

[5]. Lee GC. More data please! The evolution of orthopaedic research: commentary on an article by Cody C. Wyles, MD, et al.: “use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty”. J Bone Joint Surg Am 2019;101:118.

[6]. Fu S, Wyles CC, Osmon DR, Carvour ML, Sagheb E, Ramazanian T, et al. Automated detection of periprosthetic Joint infections and data elements using natural language processing. J Arthroplasty 2021;36:688–92. [PubMed: 32854996]

[7]. Tibbo ME, Wyles CC, Fu S, Sohn S, Lewallen DG, Berry DJ, et al. Use of natural language processing tools to identify and classify periprosthetic femur fractures. J Arthroplasty 2019;34:2216–9. [PubMed: 31416741]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

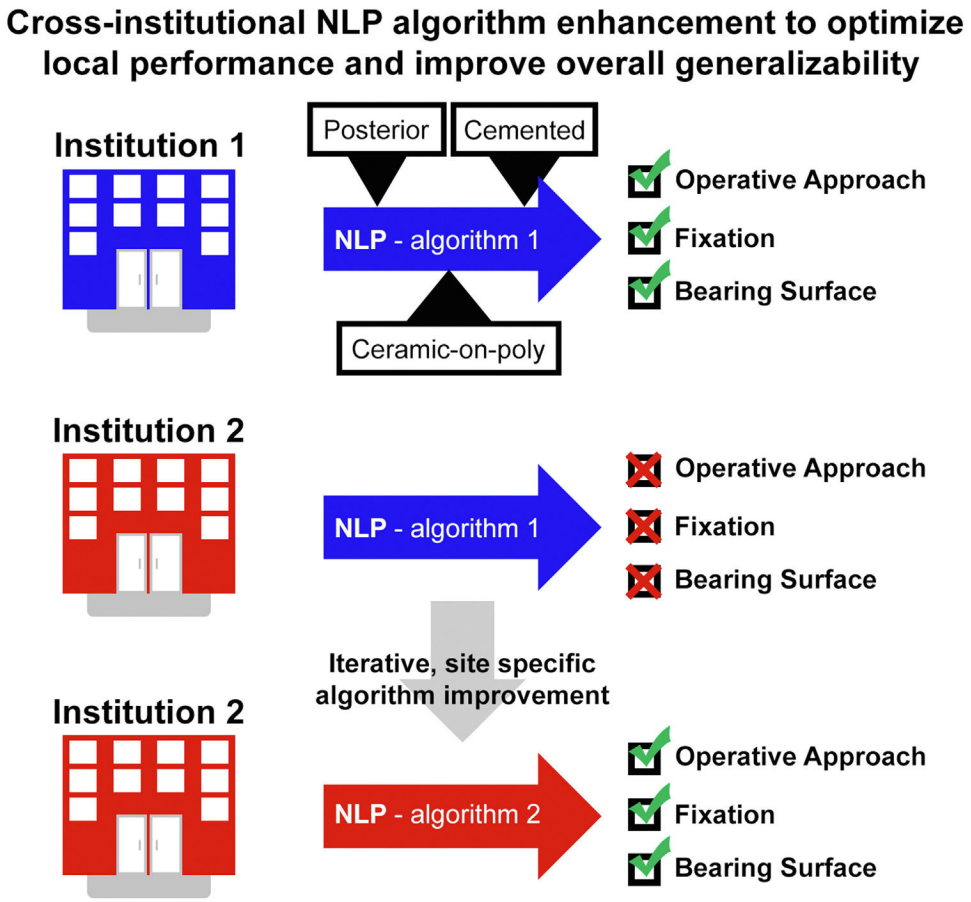

Fig. 1.

Schematic demonstrating improved capacity for local NLP use, interinstitutional sharing, and AJRR participation.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

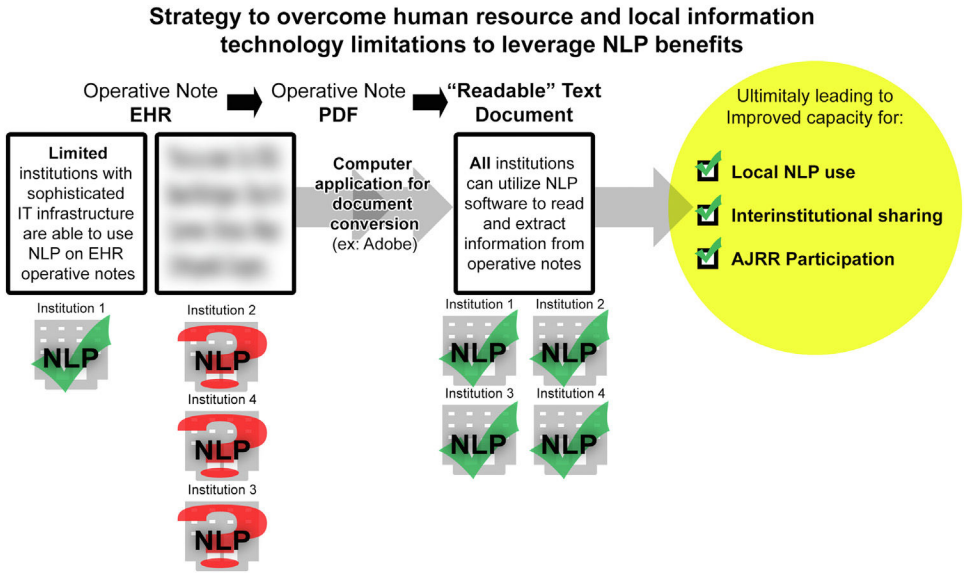

Fig. 2.

Schematic demonstrating strategy to overcome human resource and local information technology limitations to leverage NLP benefits.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Algorithm Accuracy in Tested Practice Settings.

Data Point	Mayo Clinic Rochester[a] N = 300	Mayo Clinic Health System[b] N = 180	OrthoCarolina[c] N = 39
Surgical Approach	99.2%	94.4%	100%
Fixation	90.7%	95.6%	100%
Bearing Surface	95.8%	98.0%	100%

[a] Original institution.

[b] First external validation.

[c] Second external validation.